24.2 Context-Aware Hierarchical Information-Sensing in a 6µW 90nm CMOS Voice Activity Detector

Komail Badami, Steven Lauwereins, Wannes Meert, Marian Verhelst

KU Leuven, Leuven, Belgium

The rise of always-listening sensors integrated in energy-scarce devices such as watches and remote-controls increases the need for intelligent scalable interfaces. Contemporary sensor interfaces digitize raw sensor data to extract information with energy-intensive computations, such as FFT, which is inefficient if the end goal is to only extract selective information for classification tasks, e.g. voice activity detection (VAD). Previous work shows energy gains from early data reduction through analog feature extraction [1] or embedded classification hardware [2]. However, the potential energy savings of these devices is limited as they cannot adapt to changes in the sensed information content or sensing context, such as the amount/type of acoustic background noise. In the processor design community, such adaptivity to varying operating conditions is actively researched through the concept of hierarchical computing [3]. This work integrates the concept of hierarchical operation with adaptive early data extraction and classification, towards a power- and context-aware information-extraction sensor interface. This paper specifically reports on a µW 90nm CMOS VAD, that dynamically adapts sensing resources to signal information content and context, thus only spending energy on relevant information extraction. An order of magnitude in power savings is achieved by exploiting hierarchical sensing, run-time activated/scalable analog feature extraction and tightly-integrated context-aware mixed-signal machine learning inference, enabling novel applications in area of acoustic sensing [1, 4].

Figure 1 illustrates the high-level architecture and operating paradigm. A classical, yet configurable, always-listening wake-up detector (A) operates in nW range. Upon detection of potential information, a more powerful scalable analog feature extractor and embedded mixed-signal machine learning classification block (B) are activated, operating in the µW range. These blocks extract and process a feature subset and are programmed to achieve high classification accuracy within the present operating context, as determined by the amount and type of acoustic background noise. A context-aware control register (CR) only activates the most discriminating features for the current context and configures the analog feature extractor to the desired trade-off between detection accuracy and power consumption depending on QoS and power constraints. Based on the activated features, an embedded mixed-signal decision tree (DT) classifier evaluates the signal relevance and, upon interest detection, wakes up the off-chip micro-processor (µP) (C). The µP is responsible for more advanced acoustic signal processing (e.g. keyword detection), periodic context detection, relearning of the DT in case of context change and reprogramming the CR. The outlined hierarchical activation scheme results in an elastic power consumption of the sensing chip, which dynamically scales with the amount of information present in the sensed signal. The context-awareness on the other hand enables state-of-the-art (SotA) detection accuracy across disparate operating contexts while only spending energy on extracting information-bearing data.

The configurable wake-up detector (top of Fig. 2) operates below 750nW and activates mode B if the input signal exceeds a µP-set threshold as seen at the top of Fig 3. Varying the comparator threshold controls how often the feature extractor and classifier are activated, trading-off overall accuracy vs. power consumption. The context-scalable analog feature extractor (bottom of Fig. 2) extracts the energy-content of the incoming signal in 16 Mel-spaced frequency bands between 75Hz and 5kHz, resulting in 16 individually activated analog features (*af1* - *af16*). Each band consists of an amplifier and BPF followed by a rectifier and LPF. As the DT is trained with the chip's own analog features, it automatically adapts to any process variations of the BPF characteristics. Fig. 3 shows the measured response of 4 selected analog features to a sine wave frequency sweep (bottom left) and the measured analog performance (bottom right). The DT-based mixed-signal classifier (left side of Fig. 4) can be configured to any 7-node (3-level deep) DT (or less) taking decisions on any combination of *af5* to *af12,* as they carry the highest information to power consumed ratio for VAD. The particular DT configuration and required tree reference levels ($Vref_i$) are adapted to the acoustic context and system's energy constraints by the µP. To this end, the µP periodically has access to all features (*af1*-*af16)* to detect context change and learns at run-time a new DT optimized for that new context, enabling power efficient DTs while maintaining SotA accuracy. This learning phase on the µP [6] optimizes the tree using information-gain/watt as a cost function instead of the commonly used information-gain, to identify the subset of analog features that result in the lowest power consumption for a given miss-detect/false-alarm accuracy. The configurable DT implementation consists of an analog feature selection stage, a reference comparison stage and a digital decision fusion stage. The feature selection stage maps the acoustic features (*af*) to the desired selected features (*sf*) for every decision node (Note that one *af* can map to multiple *sf*). In the comparison stage, the 7 selected features are compared to 7 reference levels set by the µP through external DACs. An invert bit selects between $sf_i > Vref_i$ or $sf_i \leq Vref_i$. The digital decision fusion stage implements the tree structure to produce a single voice detection signal waking-up the µP. The right side of Fig. 4 shows measured speech/non-speech detection accuracies for various signal to acoustic noise ratios (SANR). Audio streams with a duration of 168s, from the NOIZEUS [5] database, containing 50% voice are sent through the analog feature extraction block. Subsequently, the acoustic features *af5*-*af12* measured on the chip are used offline to train DT's on the achievable trade-off curve between speech/non-speech accuracy. Finally, one trade-off point is selected and the corresponding DT is configured on chip in the embedded classifier. Measurements (black-squares) confirm the performance of the analog feature extractor and embedded DT classifier.

Fig. 5 depicts the benefits of bringing the full hierarchical sensing system together. While every operating mode ensures a low miss-detection rate, the false-alarm rates and context-specificity are systematically decreased with the gradual wake-up of more powerful modes upon interest detection. Always-on mode A ensures low average power consumption, operating well below 1µW. Context-specific mode B does a power-efficient drastic reduction of the false alarm rate, minimizing the power-expensive start-up of the mode C which ensures that the system works across heterogeneous contexts. The power hungry µP sporadically activates to check the stability of the operating context and performs run-time embedded machine learning of a new DT in case of a context switch. Table 5 shows that this hierarchical context-aware VAD has a voice/noise accuracy of 89/85% for 12dB SANR babble noise, on par with SotA software VADs [7] yet consuming only 3.8µW on average for hybrid operation.

Figure 6 compares our hierarchical context-aware 90nm CMOS VAD chip (Fig. 7) to analog/digital/software SotA VADs. The presented VAD does pay a penalty of a larger latency in voice detection, however staying within acceptable range for natural speech applications. The worst case power consumption of the VAD chip is 6µW performing well below the current SotA. The tight integration of hierarchical context-aware analog feature extraction with on chip mixed-signal classification clearly demonstrates superior energy efficiency, while maintaining SotA accuracies on standardized speech/noise databases. The presented paradigm opens up numerous other acoustic event detection applications, ranging far beyond VAD, and can also be ported to other sensor interfaces, such as gesture recognition. This work was funded by the FWO-Flanders, the IWT SBO project SINS and an EXPERTS scholarship.

References:

[1] B. Rumberg, et al., "Hibernets: Energy-Efficient Sensor Networks Using Analog Signal Processing", *J. Emerging and Selected Topics in Circuits and Systems,* vol. 1, pp. 321-334, Sept. 2011

[2] J. Lu, et al., "A 1TOPS/W Analog Deep Machine-Learning Engine with Floating-Gate Storage in 0.13µm CMOS", *ISSCC Dig. Tech Papers*, pp. 504-506, Feb. 2014

[3] A.Wang, et al,"Heterogeneous Multi-Processing Quad-Core CPU and Dual-GPU Design for Optimal Performance, Power and Thermal Tradeoffs in 28nm Mobile Appl Processor", *ISSCC Dig. Tech Papers*, pp. 180-182, Feb. 2014

[4] A. Raychowdhury, et al., "A 2.3 nJ/Frame Voice Activity Detector Based Audio Front-End for Context-Aware SoC Applications in 32-nm CMOS", *J. Solid-State Circuits*, vol. 48, pp. 1963-1969, Aug. 2013.

[5] Y. Hu, et al., "Subjective evaluation and comparison of speech enhancement algorithms", *Speech Communication*, vol. 49, pp. 588-601, 2007

[6] S. Lauwereins, et al., "Ultra-low-power Voice-activity-detector through Context- And Resource-cost-aware Feature Selection in Decision Trees", *Int. Workshop on Machine Learning for Signal Processing*, Sept. 2014

[7] J. Kola, et al., "Voice Activity Detection," *MERIT BIEN*, pp. 1-6, 2011

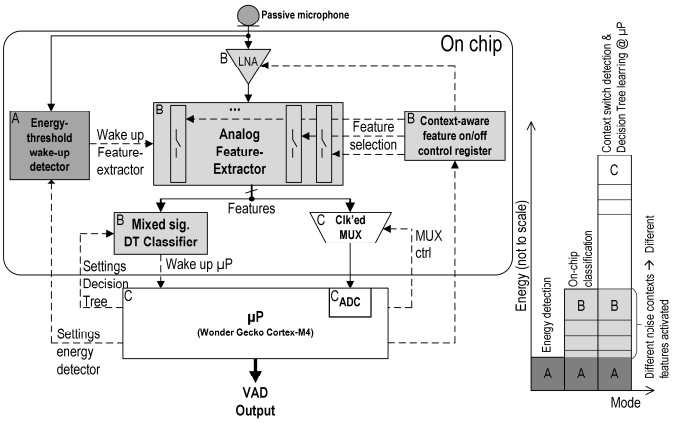


Figure 24.2.1: (left) Architectural representation of voice activity detector detailing hierarchical information extraction (right) energy consumption at different levels of hierarchy.

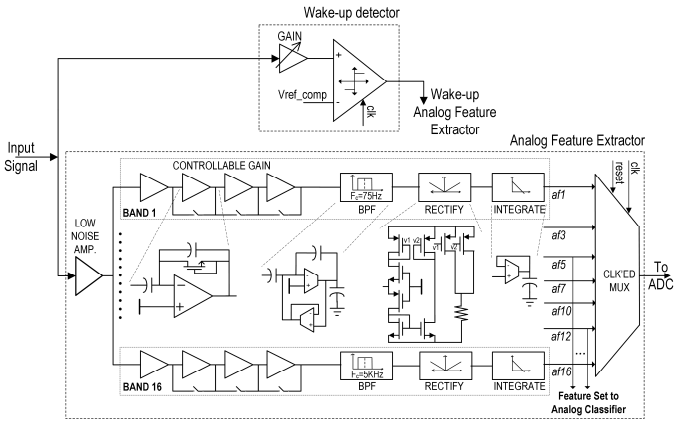
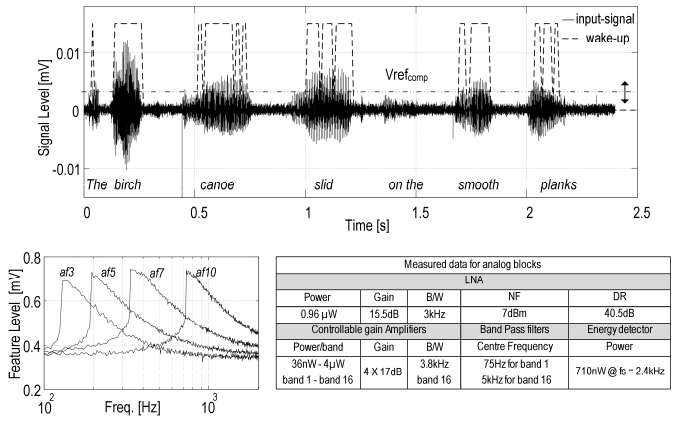


Figure 24.2.2: Schematic representation of (top) Wakeup detector (bottom) Analog feature extractor



Measured data for analog blocks				
LNA				
Power	Gain	BW	NF	DR
0.96 µW	15.5dB	3kHz	7dBm	40.5dB
Controllable gain Amplifiers			Band Pass filters	Energy detector
Power/band	Gain	BW	Centre Frequency	Power
36nW - 4µW band 1 - band 16	4 X 17dB	3.8kHz band 16	75Hz for band 1 5kHz for band 16	710nW @ fs ~ 2.4kHz

Figure 24.2.3: (top) Measured response of Wakeup to audio input (bottom left) measured band frequency response and (bottom right) measured performance summary of analog feature extraction block and energy detector

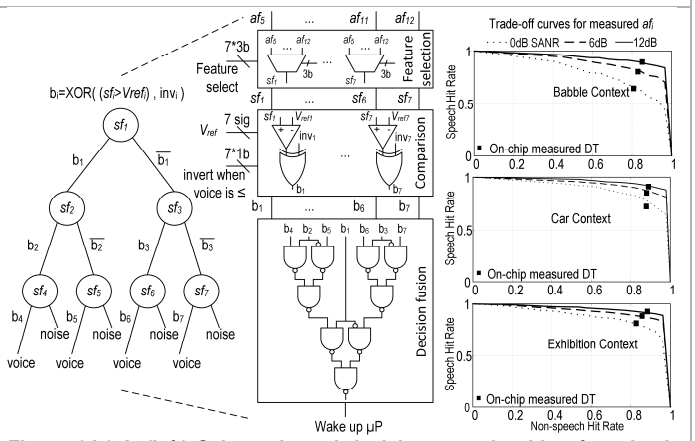
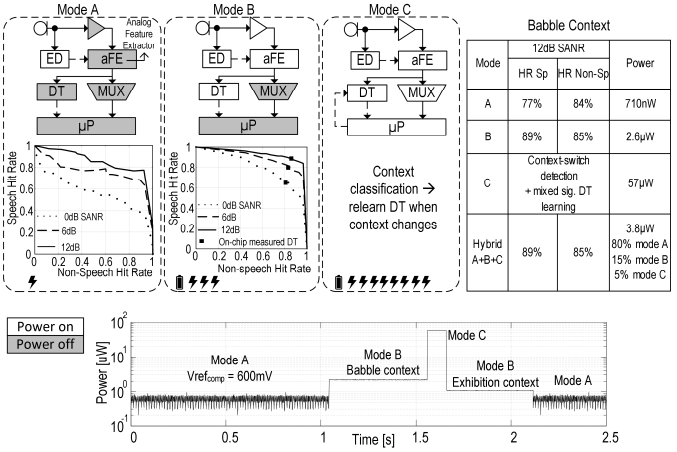


Figure 24.2.4: (left) Schematic and decision tree algorithm for mixed-signal classifier (right) Measurement results for HR speech / Non speech for different contexts.



Babble Context			
Mode	12dB SANR		Power
	HR Sp	HR Non-Sp	
A	77%	84%	710nW
B	89%	85%	2.6µW
C	Context-switch detection + mixed sig. DT learning		57µW
Hybrid A+B+C	89%	85%	3.8µW 80% mode A 15% mode B 5% mode C

Figure 24.2.5: Measured power consumption and Speech / Non Speech Hit rates for different operating modes and contexts

	This Work	[1] JETCAS '11	[4] JSSC' 13	[7]
Tech.	90nm CMOS	0.5um CMOS	32nm CMOS	Software only
Area	2mm^2	2.25mm^2	86K gates	NA
Power (feature extraction + classification)	6µW Worst case, all bands on	51µW	< 50µW	>90µW estimated [6]
Gain ncesary for passive mic.	On chip	Off chip	assumes digital mic.	NA
Feature type	Analog	Analog	Digital	Software
Classifier	On chip - Mixed Signal	Off chip - Digital	On chip - Digital	Software based
Context Aware	Yes	NA	Yes	Yes
Feature-Cost aware	Yes	NA	No	No
Latency	< 100ms	100ms	10ms	10ms
Classifier accuracy @ 12dB SNR	HR SP 89% HR Non SP 85% @ Babble 12dB SANR	90% car vs truck classification	97% Unspecified SNR / context / database	HR SP 89% HR Non SP 79% @ Babble 12dB SANR

Figure 24.2.6: Comparison to state-of-the-art.

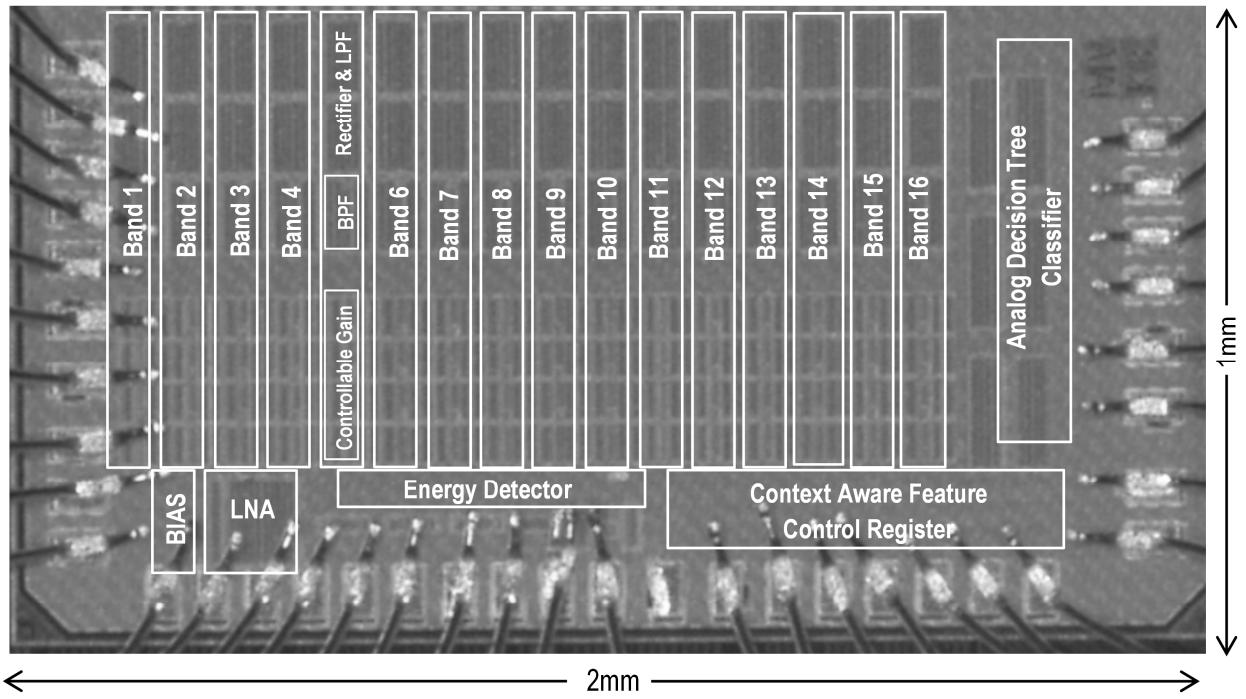


Figure 24.2.7: Chip micrograph highlighting different sections